John von Neumann Institute for Computing


# A Parallel Workflow for the Reconstruction of Molecular Surfaces

Daniele D'Agostino, Ivan Merelli, Andrea Clematis,
Luciano Milanesi, Alessandro Orro

published in

*Parallel Computing: Architectures, Algorithms and Applications* ,
C. Bischof, M. Bücker, P. Gibbon, G.R. Joubert, T. Lippert, B. Mohr,
F. Peters (Eds.),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **38**, ISBN 978-3-9810843-4-4, pp. 147-154, 2007.
Reprinted in: Advances in Parallel Computing, Volume **15**,
ISSN 0927-5452, ISBN 978-1-58603-796-3 (IOS Press), 2008.



http://www.fz-juelich.de/nic-series/volume38

# A Parallel Workflow for the Reconstruction of Molecular Surfaces

**Daniele D'Agostino**[1], **Ivan Merelli**[2], **Andrea Clematis**[1], **Luciano Milanesi**[2], **and Alessandro Orro**[1]

[1] Institute for Applied Mathematics and Information Technologies, National Research Council
Via De Marini 6, 16149 Genova, Italy
*E-mail: {dago, clematis}@ge.imati.cnr.it*

[2] Institute for Biomedical Technologies, National Research Council
Via Fratelli Cervi 93, 20090 Segrate (MI), Italy
*E-mail: {ivan.merelli, luciano.milanesi, alessandro.orro}@itb.cnr.it*

In this paper a parallel workflow for the reconstruction of molecular surfaces based on the isosurface extraction operation is proposed. The input is represented by the atomic coordinates of a molecule, the output is both its volumetric description and the isosurface that, on the basis of the user selection, corresponds to the Van der Waals, Lee & Richards or Connolly surface. The main feature of the workflow is represented by the efficient production of high resolution surfaces. This is a key aspect in Bioinformatics applications, considering that the amount of data to process may be very huge. This goal is achieved through a parallel implementation of the stages of the workflow.

## 1 Introduction

The modelling of molecular surfaces and their visualization is assuming an increasing importance in many fields of Bioinformatics. This is particularly true in the study of molecule-molecule interactions, usually referred as molecular docking, that represents one of the subject of our researches.

Superficial complementarities play a significant role to determine the possible binds between pairs of molecules[1,2]. Mechanisms such as enzyme catalysis and recognition of signals by specific binding sites and docking in fact depend on morphological characteristics. It is clear that a correct superficial description plays a crucial role for the definition of a valid docking. Usually a molecule is represented through the set of its 3D atomic coordinates. This is for example the format adopted by the Protein Data Bank (PDB)[3], one of the most important repositories. Such format well suits structural analysis operations, but not those based on the molecular shape. This is the reason why other representations, that can be derived from this one, have to be taken into account.

In particular for molecular docking the Van der Waals[4], the Lee & Richards[5] and the Connolly[6] surfaces are the most important ones. All of them are computed by modelling the atoms with spheres, and the result is the production of polygonal (mostly triangular) meshes, but they differ for the considered atomic radii and the possibility to close some small superficial cavities.

Different algorithms were proposed in the literature to compute such surfaces, and many implementations are available. Most of them rely on a pure geometrical analysis of the atoms' positions and on the estimation of their volumes. One example is MSMS[7],

that is probably the most used tool. The main characteristic of this approach is its computational efficiency, but disregarding the volumetric description some information about binding sites buried under the surface is lost. This is a major drawback for the docking analysis. Other algorithms rely on the volumetric description of the molecule using the isosurface extraction operation. This approach permits an accurate modelling of its inside structures, but is more costly than the previous one. The most representative tool in this case is GRASP[8]. Both these tools have the drawback than they run out of memory for large surfaces, therefore they are able to provide only limited resolution surface representations. This may be an important issue considering the need to compute the buried surface upon molecular formation, therefore MSMS and GRASP are not perfectly suited for an in-depth analysis of internal functional sites.

In this paper a parallel workflow for the reconstruction of molecular surfaces based on the isosurface extraction operation is proposed. Two are the main advantages, the performace and the surface quality. With regards to the performance, the tests showed that also using the sequential implementation of the workflow it is possible to compute the molecular surfaces faster than MSMS and GRASP.

Moreover the use of the simplification operation permits to produce high quality surfaces with different levels of detail. The purpose of this operation is to preserve the morphological information in correspondence of irregular zones, that are the most interesting ones, with great accuracy, while reducing the number of triangles in the other parts. The resulting simplified surface is therefore made up by less triangles than the original one, but it preserves all the important features of the molecule. This aspect is of particular importance for molecular docking, that has a cost proportional to the size of the surfaces to process. In this case the parallel implementation of the workflow is fundamental, because it permits to obtain fast results also for surfaces made up by several million triangles, that otherwise cannot be processed.

The paper is organized as follows. In Section 2 the structure of the workflow and the parallel implementations of the various stages is described. Section 3 presents the experimental results, while conclusions and future work are outlined in Section 4.

## 2 The Workflow Design and Implementation

The architecture of the workflow is represented in Fig. 1. The main input of the system is a PDB file containing the atomic coordinates of the atoms that form a molecule, while the output is represented by both the volumetric description of the molecule and the isosurface corresponding to one of the three aforementioned surfaces. The workflow is made up by five operations: *Grid Generation*, *Connolly Correction*, *Median Filter*, *Isosurface Extraction* and *Simplification*.

*Grid Generation*
The first operation is the generation of the three-dimensional grid containing the volumetric data representing the molecule. The size of the grid is determined on the basis of the coordinates of the atoms and on the required sampling step. Typical step values are chosen between 0.7 and 0.1 Å, according to the desired level of resolution. Low step values correspond to dense grids and high resolution surfaces, and vice versa.

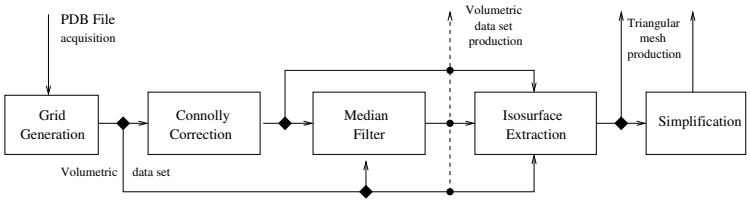



Figure 1. The stages of the parallel workflow for the reconstruction of molecular surfaces. The results are both the volumetric data set, produced before the execution of the isosurface extraction, and the triangular mesh.

Atoms are modeled in the grid with spheres having different radii. The points within a sphere assume negative values, that increase gradually from the centre to the hull, where they become positive.

Two are the possible different partitioning strategies for the parallelization of this operation. The first one is the even subdivision of the grid along the Z axis, the other one is a partitioning driven by the spatial subdivision of the atoms. Both the strategies present pros and cons, therefore they were experimentally evaluated.

The latter strategy permits to achieve the best performance considering only the execution of this operation, because the computational cost of each chunk of the grid is proportional to the number of atoms a process has to model. However this solution has the drawback to worsen considerably the performance of the whole pipeline. This is because the partitioning is not suitable for the following two operations, therefore there would be the need to perform at least one rebalancing step.

On the contrary, the former strategy permits to exploit the same partitioning for the three operations. In fact the resulting unbalancing of this stage of the workflow is less costly than the data exchange that otherwise is necessary, therefore it was decided to adopt it. Moreover this solution permits processes to keep the data in main memory until the isosurface extraction operation, if there is a sufficient amount of aggregate memory.

*Connolly Correction*

The second operation is performed only when the Connolly surface is required. This surface consists of the border of the molecule that could be accessed by a probe sphere representing the solvent. When this sphere rolls around a pair of atoms closer than its radius, it traces out a saddle-shaped toroidal patch of reentrant surface. If the probe sphere is tangent to three atoms, the result is a concave patch of reentrant surface. The main difference with respect to the Van der Waals and the Lee & Richards surfaces is that these patches close the superficial small cavities.

The Connolly Correction operation consists in changing the values of the points of the volume that become internal (and so with a negative value) considering these new patches. It is performed in two steps, the identification of the pairs of close atoms and the modification of the values of the points in the neighbourhood of these pairs.

The first step requires to compare the distances among all the pairs of atoms and the solvent radii. It is implemented by subdividing all the possible pairs among the processes, with the exchange of the results at the end.

The second step is implemented using the data parallel paradigm. Also in this case the

considerations made describing the parallelization strategy of the previous operation holds true, therefore its partitioning is exploited .

It is worthwhile to note that, while in the previous operation the unbalancing produced by the uneven spatial distribution of the atoms is negligible, due to their limited number, here it may be required to model the patches of more than one million pairs, therefore the unbalancing may be more relevant.

*Median Filter*
After the creation of the volumetric data set a median filter can be applied in order to produce a smoother isosurface. This operation is useful when medium-high step values are used, otherwise it may be disregarded.

The data parallel paradigm is the suitable choice also for implementing this operation. In this case in fact the amount of work is exactly proportional to the number of points to process, therefore the even partitioning of the volume guarantees high performance.

Note that the parallel processes need to exchange border regions in order to correctly compute the values of the points on the boundary of the parts. This represents a negligible overhead, except for small grids.

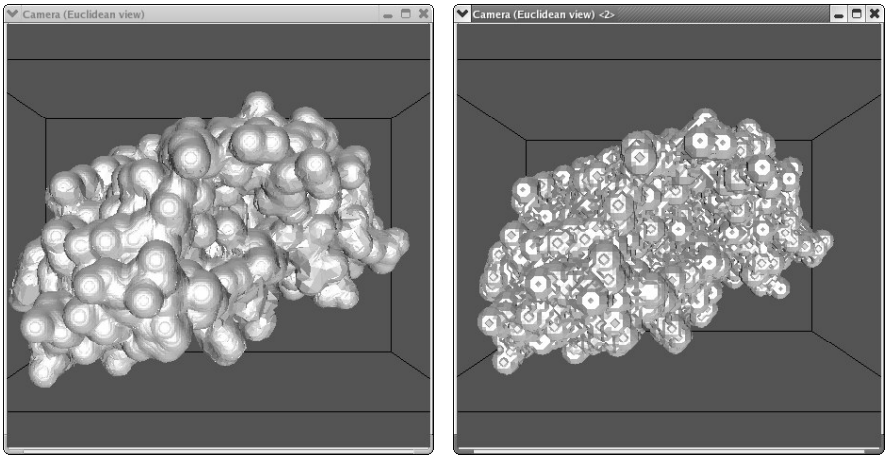
After the last of these three operations the volumetric data set is written on the hard disk, in order to make it available for further analysis operations.

*Isosurface Extraction*
The fourth operation is the extraction of the isosurface representing the molecular shape. The Connolly surface is extracted considering the isovalue 0 if the Connolly Correction is performed. Otherwise the Van der Waals and the Lee & Richards surfaces are extracted considering, respectively, the isovalues 0 and 1.4.

Two parallel versions of the Marching Cubes algorithm[9], the mostly used algorithm to extract isosurfaces from volumetric data sets, were implemented. The first one is based on the farm on demand paradigm, and provides high performance figures when the simplification operation is not performed. Otherwise the second one[10], based on a data parallel approach with an intermediate step for balancing the triangles among the parallel processes, is used. The two versions are respectively called *Farm on Demand* (FD) and *Load Balanced on Active Cells* (LBAC). A comparison of these two algorithms is discussed in[11].

*Simplification*
The Marching cubes algorithm has the characteristics of producing a large number of small planar triangles. With the simplification operation it is possible to obtain a smaller but equivalent surface by merging them. This operation is optional. It may be executed for example if the size of the isosurface overcomes a given threshold. The result is characterized by the non-uniform level of detail, because the most irregular zones are represented using more triangles than the regular ones.

If the objective is only the reduction of triangles a bigger step for Grid Generation may be used. However in this case the resulting mesh has an uniform coarse grain level of detail. In Fig. 2 two Connolly surfaces of the same molecule made up by about 310,000 triangles are presented. If the left one, obtained using a finer grid and the simplification operation with a simplification percentage of 40%, is compared with the right one, obtained using a coarser grid, it appears that the best result is obtained using the simplification.




(a) The simplified surface (b) The surface obtained using a coarser grid

Figure 2. A comparison between two different representations of the Connolly surface for the molecule of the bovine insulin "2INS". Both the images are made up by about 310,000 triangles, but the left one is obtained using a finer grid and the simplification operation with a percentage of 40%, the right one using a coarser grid.

Several simplification algorithms were proposed in the literature. Here the Garland-Heckbert algorithm[12] was chosen because it produces high quality results. A parallel version of it was developed using the data parallel approach with a coordinator process that globally selects the triangles to merge, in order to minimize the introduced representation error[13]. The boundaries of the various parts are preserved, therefore the mesh resulting after the merge of the partitions does not present any crack.

It is to note that the simplification algorithm works in core. This means that the sequential version is not able to process large meshes. The parallel version, instead, is able to process arbitrary meshes if a sufficient amount of aggregate memory is available. This aspect represents a further advantages of the parallel workflow, beside the efficiency production of the results.

## 3 Experimental Results

Experimental results were collected considering the execution of all the stages of the workflow, in order to analyze the contribution of each operation to the global execution time.

A Linux-based Beowulf Cluster of 16 PCs equipped with 2.66 GHz Pentium IV processor, 1 GB of Ram and two EIDE 80 GB disks in RAID 0 was used. The nodes are linked using a dedicated switched Gigabit network, and they share a PVFS file system.

Three molecules of the Protein Data Bank repository, chosen on the basis of their size, were considered. The smallest one is the structure of DES-PHE B1 bovine insulin, identified as 2INS and made up by 778 atoms, followed by the crystal structure of a wild type human estrogen receptor, identified as 1G50 and made up by 5,884 atoms, and by the crystal structure of the large ribosomal subunit from Deinococcus Radiodurans, identified as 1NKW and made up by 65,300 atoms.

| Molecule | Atoms | Pairs | Grid | Grid file | Triangles | Mesh file |
|---|---|---|---|---|---|---|
| 2INS | 778 | 14870 | 160x114x150 | 5.2 MB | 270568 | 4.7 MB |
| 1G50 | 5884 | 121039 | 389x294x265 | 57.8 MB | 1938128 | 33.3 MB |
| 1NKW | 65300 | 1280539 | 687x833x723 | 789.2 MB | 21813684 | 374.4 MB |

Table 1. This table summarizes the characteristics of the three considered molecules. The number of the pairs of atoms considered for the Connolly Correction is denoted with "Pairs". The size of the files containing the volumetric data and the isosurfaces is computed considering a binary format without any header. In particular the meshes are represented through the coordinates of the vertices and the triangle/vertex incidence relation.

A step of 0.3 Å was considered for the Grid Generation operation, because it permits to create also for 1NKW a grid that fits in main memory using only one node of the cluster.

The characteristics of the molecules, of the resulting volumetric data sets and the Connolly surfaces are shown in Table 1, while Table 2 shows the performance of the sequential and the parallel implementations of the workflow.

At first the sequential times were compared with those of MSMS. In particular MSMS corresponds to the execution of the Grid Generation, the Connolly Correction and the Isosurface Extraction operations. The output of the mesh was disregarded because MSMS stores it in two plain ASCII files, while here binary files are produced. It appears that the performance is the same for 2INS, while MSMS takes 2 second more for 1G50 and it is not able to produce any result for 1NKW because of the large number of atoms to model. Note furthermore that with the PCs used MSMS cannot produce meshes larger than 2 million triangles.

Considering the parallel workflow it can be seen that the efficiency is of about 0.5. This is an important result considering all the issues related to the parallelization of the workflow operations.

In particular it can be seen that the incidence of the possible unbalancing of atoms and pairs resulting from the even partitioning of the grid is less relevant when a large number of

| | 2INS | | | 1G50 | | | 1NKW | | |
|---|---|---|---|---|---|---|---|---|---|
| Nodes | 1 (sec). | 8 | 16 | 1 (sec). | 8 | 16 | 1 (sec). | 8 | 16 |
| Grid G. | 0.2 | 5.3 | 6.2 | 1.5 | 5.0 | 7.5 | 19.8 | 6.6 | 11.6 |
| Connolly C. | 0.8 | 3.0 | 4.8 | 7.4 | 3.5 | 5.3 | 215.8 | 5.2 | 8.1 |
| M. Filter | 0.4 | 6.3 | 8.1 | 4.2 | 7.9 | 15.3 | 60.0 | 8.0 | 15.8 |
| Grid Out. | 1.3 | 3.3 | 3.1 | 3.4 | 3.1 | 2.4 | 27.1 | 2.5 | 2.4 |
| Isoextr. | 2.2 | 2.5 | 3.8 | 7.2 | 2.6 | 4.7 | 70.5 | 3.4 | 5.6 |
| Simpl. | 6.7 | 5.7 | 11.8 | 48.6 | 4.8 | 8.5 | (571.3) | N.A. | 8.3 |
| Iso. Out. | 0.9 | 8.0 | 12.1 | 7.2 | 8.0 | 14.4 | (112.5) | N.A. | 13 |
| **Tot.** | **12.5** | **4.6** | **7.2** | **79.5** | **4.5** | **7.6** | **(1077)** | **N.A.** | **8.5** |

Table 2. This table presents the times, in seconds, for executing the sequential implementation of the workflow (denoted as "1"), and the speed up values of the parallel version using 8 and 16 nodes. The output of the volumetric data set is indicated with "Grid Out.", and the output of the Connolly surface with "Iso. Out.". As regards the Simplification operation the production of a mesh made up by half of the original triangles is required. With the cluster used is not possible to simplify meshes larger than 2 million triangles with only one node. For this reason an estimation of the sequential execution times for 1NKW is provided in brackets.

them have to be modeled. Furthermore this partitioning results in high speed up values for the Median Filter operation, except for 2INS. In this case in fact the overhead represented by the time to exchange the border regions among the workers is comparable to the time to process the data.

The production of the volumetric data has to be done on a shared file system, because they will be used for the isosurface extraction operation. Despite the use of PVFS, that is one of the most performant open parallel file systems, this step achieves low performance when compared with the use of a local disk. On the contrary the part of the isosurface extracted by each process is stored on the local disk of the node on which it is running, together with additional data that will be used for the efficient sewing of these chunks in a unique surface[10]. It is worthwhile to note that these additional data represent an overhead with respect to the sequential case, and furthermore their size is proportional to the number of triangles and parts to merge. For this reason the performance is slight lower for increasing values of the parallelism degree and of the isosurface size.

As said before, the execution of the parallel simplification operation implies the use of the LBAC algorithm. This algorithm presents low performance with respect to FD, but considering also the execution of the simplification it permits to achieve the highest performance[11].

Note that the parallel implementation of the workflow permits the exploitation of the aggregate memory of the cluster. This is of particular importance when considering larger grids, because in these cases the sequential workflow needs to use out-of-core techniques. For example if 1G50 with a sampling step of 0.1 Å is considered, a grid with a size of 1.5 GB (1165x884x797) has to be processed. In this case the parallel workflow achieves a speed up of 12 using 16 nodes.

Disregarding the efficiency aspects, the use of the aggregate memory has the further advantage of making possible the simplification of large meshes in order to produce smaller results with different levels of detail. With a single PC of the cluster used in fact it is not possible to simplify meshes larger than 2 million triangles, while using the cluster meshes of more than 20 million triangles as 1NKW can efficiently be processed. It is estimated that, even if it is possible to process 1NKW with a single node, the execution time should be about 18 minutes, while the parallel implementation is able to produce the result in 2 minutes.

## 4 Conclusions and Future Work

Surfaces play a fundamental role in protein functions, because chemical-physical actions are driven by their mechanic and electrostatic properties, with scarce influence of the internal structures. In this paper a method for describing molecular surfaces in detail is presented. The aim is to develop a high performance docking screening system, that represents an innovative high performance approach to the protein-protein interaction study.

The key aspects of the workflow are the efficiency and the quality of the results. These aspects are strictly coupled, in particular considering that high resolution representations of molecules is of fundamental importance for the effectiveness of the following analysis operations, but their modelling is a costly process.

As regards the performance, the parallel workflow is able to produce efficiently molecular surfaces also considering very fine grain resolutions. In particular the use of the sim-

plification operation permits an effective reduction of the surface size, preserving however a high level of detail for the irregular zones, that are the most interesting ones.

The next objective of the research is the exploitation of these surfaces in order to determine the possible docking between pairs of molecules. Also in this case the performance represents an important issue, because of the large amount of data to take into exams.

## Acknowledgements

This work has been supported by the regional program of innovative actions PRAI-FESR Liguria.

## References

1. K. Kinoshita, and H. Nakamura, *Identification of the ligand binding sites on the molecular surface of proteins*, Protein Science, **14**, 711–718, (2005).
2. T. A. Binkowski, L. Adamian, and J. Liang, *Inferring functional relationships of proteins from local sequence and spatial surface patterns*, J. Mol. Biol., **332**, 505–526, (2003).
3. H. M. Berman, T. N. Bhat, P. E. Bourne, Z. Feng, G. Gilliland, H. Weissig, and J. Westbrook, *The Protein Data Bank and the challenge of structural genomics*, Nature Structural Biology, **7**, 957–959, (2000).
4. M. L. Connolly, *Solvent-accessible surfaces of proteins and nucleicacids*, Science, **221**, 709–713, (1983).
5. B. Lee, and F. M. Richards, *The Interpretation of Protein Structures: Estimation of Static Accessibility*, J. Mol. Biol., **55**, 379–400, (1971).
6. M. L. Connolly, *The molecular surface package*, J.Mol.Graphics, **11(2)**, 139–141, (1993).
7. M. F. Sanner, A. J. Olson, and J. Spehner, *Fast and Robust Computation of Molecular Surfaces*, in: Proc. 11th ACM Symp. Comp. Geometry, C6–C7, (1995).
8. A. Nicholls, K. A. Sharp, and B. Honig, *Protein Folding and Association: Insights From the Interfacial and Thermodynamic Properties of Hydrocarbons*, Proteins: Stuc., Func. and Genet., **11**, 281–296, (1991).
9. W. E. Lorensen, and H. E. Cline, *Marching cubes: A high resolution 3-D surface construction algorithm*, Computer Graphics, **21(3)**, 163–169, (1987).
10. A. Clematis, D. D'Agostino, and V. Gianuzzi, *An Online Parallel Algorithm for Remote Visualization of Isosurfaces*, Proc. 10th EuroPVM/MPI Conference, LNCS No. 2840, 160–169 (2003).
11. A. Clematis, D. D'Agostino, and V. Gianuzzi, *Load Balancing and Computing Strategies in Pipeline Optimization for Parallel Visualization of 3D Irregular Meshes*, Proc. 12th Euro PVM/MPI Conference, LNCS No. 3666, 457–466 (2005).
12. M. Garland, and P. S. Heckbert, *Surface Simplification Using Quadric Error Metrics*, Computer Graphics, **31**, 209–216, (1997).
13. A. Clematis, D. D'Agostino, V. Gianuzzi, and M. Mancini, *Parallel Decimation of 3D Meshes for Efficent Web based Isosurface Extraction*, Parallel Computing: Software Technology, Algorithms, Architectures & Applications, Advances in Parallel Computing series, **13**, 159–166, (2004).